Spatial gradient in value representation along the medial prefrontal cortex reflects individual differences in prosociality

Sul, Sunhae ; Tobler, Philippe N ; Hein, Grit ; Leiberg, Susanne ; Jung, Daehyun ; Fehr, Ernst ; Kim, Hackjin

Abstract: Despite the importance of valuing another person's welfare for prosocial behavior, currently we have only a limited understanding of how these values are represented in the brain and, more importantly, how they give rise to individual variability in prosociality. In the present study, participants underwent functional magnetic resonance imaging while performing a prosocial learning task in which they could choose to benefit themselves and/or another person. Choice behavior indicated that participants valued the welfare of another person, although less so than they valued their own welfare. Neural data revealed a spatial gradient in activity within the medial prefrontal cortex (MPFC), such that ventral parts predominantly represented self-regarding values and dorsal parts predominantly represented other-regarding values. Importantly, compared with selfish individuals, prosocial individuals showed a more gradual transition from self-regarding to other-regarding value signals in the MPFC and stronger MPFC–striatum coupling when they made choices for another person rather than for themselves. The present study provides evidence of neural markers reflecting individual differences in human prosociality.

DOI: https://doi.org/10.1073/pnas.1423895112

Originally published at:
Sul, Sunhae; Tobler, Philippe N; Hein, Grit; Leiberg, Susanne; Jung, Daehyun; Fehr, Ernst; Kim, Hackjin (2015). Spatial gradient in value representation along the medial prefrontal cortex reflects individual differences in prosociality. Proceedings of the National Academy of Sciences of the United States of America, 112(25):7851-7856.
DOI: https://doi.org/10.1073/pnas.1423895112

BIOLOGICAL SCIENCES/Psychological and Cognitive Sciences

Spatial gradient in value representation along the medial prefrontal cortex reflects individual differences in prosociality

Sunhae Sul[a, e], Philippe N. Tobler[b], Grit Hein[b], Susanne Leiberg[b], Daehyun Jung[c], Ernst Fehr[b], Hackjin Kim[a, d]

[a]Department of Psychology, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 136-701, Republic of Korea

[b]Department of Economics, University of Zurich, Blümlisalpstrasse 10, CH-8006, Zurich, Switzerland

[c]Department of Brain and Cognitive Engineering, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 136-701, Republic of Korea

[d] Correspondence concerning this article should be addressed to Hackjin Kim, Department of Psychology, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 136-701, Republic of Korea. Email: hackjinkim@korea.ac.kr

[e] Present address: Sunhae Sul, Ph.D., Department of Psychological and Brain Sciences, Dartmouth College, 6207 Moore Hall, Hanover, NH 03755, USA. (sunhae.sul@dartmouth.edu)

Keywords: medial prefrontal cortex | striatum | anterior insula | computational model | reinforcement learning

Abstract

Despite the importance of valuing another person's welfare for prosocial behavior, currently we have only a limited understanding of how these values are represented in the brain, and more importantly, how they give rise to individual variability in prosociality. In the present study, participants underwent functional magnetic resonance imaging while they performed a prosocial learning task in which they could choose to benefit themselves and/or another person. Choice behavior indicated that participants valued the welfare of another person, although less so than they valued their own welfare. Neural data revealed a spatial gradient in activity within the medial prefrontal cortex (MPFC), such that ventral parts predominantly represented self-regarding values whereas dorsal parts predominantly represented other-regarding values. Importantly, compared to selfish individuals, prosocial individuals showed a more gradual transition from self- to other-regarding value signals in the MPFC and stronger MPFC-striatum coupling when they made choices for another person rather than for themselves. The present study provides evidence of neural markers reflecting individual differences in human prosociality.

Significance Statement

How do selfish and prosocial brains function differently with regard to valuing the welfare of others? The present study addresses this question by combining neuroimaging, computational modeling, and an instrumental conditioning paradigm. Contrary to the conventional notion of the dorsal medial prefrontal cortex (MPFC) implicated in mentalization, we found that it was selfish individuals who showed greater spatial segregation between ventral and dorsal MPFC, which encoded self- and other-regarding values, respectively. Prosocial individuals, on the other hand, were characterized by overlapping representations of self-value and other-value in the ventral MPFC and also by stronger functional coupling between MPFC and striatum while representing and updating the value of other-regarding choices. These findings provide rigorous scientific evidence of neural markers reflecting individual differences in human prosociality.

\body

Ranging from a small act of kindness in daily life to self-sacrificing altruism under life-threatening situations, we often observe large individual differences in valuing another person's welfare. This differential valuation process seems to be the key to understanding various human prosocial behaviors, which are fundamental to the sustainability of human society (1). Yet, the underlying neural mechanisms and their relationship to individual differences in prosociality remain unclear.

Perhaps the most powerful way of assessing how a future outcome is valued is to use an instrumental learning paradigm that examines whether the occurrence of a response increases when it is followed by that outcome (2). The mechanisms underlying this type of learning have been described more formally with a computational model, known as the temporal difference (TD) learning model (3-5), which has been used successfully to reveal the neuroanatomical substrates of subjective valuation (3,4,6). Previous research has further refined the neurobiological model of reinforcement learning by emphasizing the specific roles played by the medial frontal cortex and the striatum: the medial frontal cortex computes the value of the chosen action while the striatum processes reward prediction errors during reinforcement learning (4, 6-9).

Unlike our current understanding of the valuation process for self-regarding choices (3, 6-12), it is much less clear whether learning can also be driven by other-regarding values, and whether this other-regarding valuation relies on the same mechanisms of reinforcement learning as those used for self. Moreover, despite rapidly

accumulating research on reward processing in social domains (13-19), the question remains of how the neural representation of self- vs. other-regarding values is related to individual differences in altruistic behavior.

In the present research, we designed a novel version of an instrumental learning task (i.e., a prosocial learning task), to assess behavioral and neural processes associated with self- and other-regarding valuation in a comparable, principled way. In the prosocial learning task, participants chose between two alternatives to achieve a higher probability of benefitting either themselves and/or another person in the form of reducing the duration of exposure to unpleasantly loud noise. Thirty pairs of healthy right-handed female college students participated in the study. The scanned participant of each pair performed the prosocial learning task (**Figure 1**). In each trial of the task, participants were presented with two options and had to choose one of them. In different conditions, the two options were represented by specific fractal images and associated with points only for the participant in the scanner (SELF condition), for both participants (BOTH condition), or only for the participant outside the scanner (OTHER condition). One of the two options always had a higher probability (70%) of yielding points than the other (30%). By trial and error the participants in the scanner would learn about these probabilities and subsequently choose the option they preferred. Participants were told that they would be exposed to unpleasant noise for five minutes after the task, and that the points earned in the task would be used to reduce the duration of the noise for themselves and/or the paired participant outside the scanner.

We predicted that if participants valued others' welfare, the other-regarding outcome (i.e., points earned to reduce the duration of aversive noise for the other) would increase their performance above chance level, such that they would earn more points for the other participant than if they chose randomly. In line with the idea that avoidance of punishment is reinforcing and has been shown to activate similar brain regions as reward learning (3), the points earned in the task, which could be used later, just like money, were presumed to have appetitive motivational value. Regarding neural representation of the chosen value, therefore, we expected a spatial segregation within the MPFC in computing self- and other-regarding values, consistent with previous studies showing that the ventral and dorsal parts of the MPFC are involved in self- and other-regarding processes, respectively (19-25). More importantly, we hypothesized that the degree of spatial segregation would provide a neural index of the individual propensity to help others. Given that the positive subjective valuation of others' welfare can lead to prosocial decisions (17, 26-29), we expected that decreased segregation would be associated with greater prosociality. In addition, we examined whether and how corticostriatal communications contribute to individual differences in representing and updating self- and other-regarding values.

Behavioral results

We tested whether participants valued another person's welfare at all, and found that they did. In particular, the proportions of choosing the high-reward probability option

in the OTHER condition were significantly higher than chance level (0.5), $t(25) = 2.68$, $p < 0.05$ (**Figure 2A**; see **SI Appendix**, **Figure S1** for trial-by-trial learning curves). While this finding clearly indicates that participants did learn to help others even when there was nothing to be gained for them, there were considerable individual differences in the propensity to help. Some individuals showed equal preference for the high probability option in the SELF and the OTHER conditions whereas other individuals showed such a preference only in the SELF condition. Due to this individual variability, preference for the high probability option was weaker, on average in the OTHER condition compared to the SELF (ps < 0.05) or the BOTH (ps < 0.05) conditions [**Figure 2A**; main effect of condition: $F(2, 50) = 5.97$, $p < 0.01$; pairwise comparisons for SELF vs. BOTH: ns].

fMRI Results

Spatial gradient within MPFC for self- vs. other-regarding value computation. We hypothesized that the ventral and dorsal subregions of the MPFC would be involved in computing self- and other-regarding values, respectively. A minimal requirement to support this hypothesis was that the MPFC as a whole would be associated with choice values in all conditions. We therefore conducted a parametric modulation analysis using subject-specific value parameters estimated by the advantage learning model (see **Materials and Methods**) and found that the MPFC (x = 4, y = 52, z = 8 mm, $Z = 3.73$) was engaged in computing the value of the chosen option at the time of stimulus presentation across all three conditions (**SI Appendix, Table S1** and **Figure S3A**). Next, we ran the same parametric modulation analysis separately for each condition. These

analyses revealed that the chosen value-related MPFC activation clusters in the SELF and the OTHER conditions were respectively located somewhat more ventrally and more dorsally than the cluster in the BOTH condition (**SI Appendix, Table S1** and **Figure S3B**), consistent with our prediction of spatial specificity.

For a more quantitative examination of this spatial segregation within the MPFC, we defined five regions of interest (ROIs) along the ventral-dorsal midline axis. Specifically, we obtained a sagittal view of the statistical parametric map from the aforementioned parametric modulation analysis with a lenient threshold of $p < 0.05$ uncorrected and then selected five equally spaced coordinates spanning the ventral to dorsal extent of the MPFC (**Figure 3A**), similarly to previous research (30). The parameter estimates of the neural activation associated with chosen value at the time of stimulus presentation were extracted from the ROIs for each individual. A 2 (condition: SELF, OTHER) × 5 (ROI locations) repeated-measures ANOVA showed a clear spatial distinction between VMPC and DMPFC in computing values for self- vs. other-regarding choices [interaction of condition with ROI locations: $F(4, 100) = 4.49$, $p < 0.005$] (**Figure 3B**). The value signal for SELF was stronger in more ventral ROIs, whereas the value signal for OTHER was stronger in more dorsal ROIs within the MPFC. Value signals for the BOTH condition were dominant in intermediate ROIs, and the interaction effect remained significant when we included the BOTH condition in the analysis [$F(8, 200) = 2.28$, $p < 0.05$]. In line with the gradient hypothesis, the linear effect of ROI location on condition was also significant [$F(1, 25) = 10.13$, $p < 0.005$]. More specifically,

the strength of the value signal for self-regarding choices linearly decreased from VMPFC to DMPFC [$F(1, 25) = 4.81$, $p < 0.05$] whereas the value signal for other-regarding choices showed the opposite trend [$F(1, 25) = 3.10$, $p = 0.09$].

Self-other distinction within the MPFC reflects propensity to help others in the prosocial learning task. We expected that the spatial pattern of value representations within the MPFC would reliably track individual variability in (prosocial) propensity to help, as measured by choice behavior in the prosocial learning task. To capture individual differences in a categorical manner, we formed two groups (prosocial and selfish) based on the parameters estimated by the advantage learning model (see **Materials and Methods**). **Figure 2B** illustrates the behavioral characteristics of the prosocial and selfish individuals in terms of their propensity to choose high reward probability options across conditions (see **SI Appendix, Figure S1** for the learning curves). Group membership (prosocial and selfish) interacted with condition (SELF, BOTH, and OTHER) with respect to the proportions of high probability choice [$F(2, 46) = 11.78$, $p < 0.001$]. Individuals in the prosocial group showed a smaller difference between the SELF and OTHER condition than individuals in the selfish group (see **SI Appendix, Figures S2** and **S4** for the validation of the categorization).

To better characterize the effects of prosociality on value representation, we conducted a mixed ANOVA with reward-type and ROI location as within-subject factors and group membership as a between-subjects factor. We found a significant three-way interaction [$F(4, 92) = 3.10$, $p < 0.05$], such that the spatial distinction of self- and other-

regarding value representations was stronger in the selfish group than the prosocial group (**Figures 3C** and **3D**). Analyses performed separately for the prosocial and selfish groups further supported this finding: The spatial separation of self- and other-regarding value signals was only prominent among selfish individuals, [$F(4, 36) = 5.37$, $p < 0.005$] but not among prosocial individuals [$F(4, 56) < 1$, ns]. We quantified the degree of spatial separation within the MPFC by fitting linear functions to the self- and other-regarding value signals along the ventral-dorsal axis for each individual. A between-groups comparison of the linear slopes fitted to the spatial gradient revealed that slopes were steeper in selfish vs. prosocial individuals [$F(1, 23) = 6.72$, $p < 0.05$] (**Figure 3E**), consistent with a greater separation of self- and other-regarding values in selfish compared to prosocial individuals. The spatial gradient revealed that the difference between selfish and prosocial groups mainly arose in the OTHER condition [$F(1, 23) = 4.015$, $p = 0.057$], that is, the other-regarding value signal was stronger in the DMPFC than the VMPFC only in selfish individuals. Such a group difference was not observed in the SELF condition [$F(1 ,23) < 1$, n.s.], where the self-regarding value signal was stronger in the VMPFC than the DMPFC for both groups. Control analyses showed that the extent of the MPFC activation was not merely associated with performance level (see **SI Appendix** for additional analyses addressing alternative explanations).

Furthermore, the spatial gradient tracked the choices in the prosocial learning task: the slopes of the spatial gradient for self- and other-regarding values correlated negatively with the average proportion of choosing high reward probability options in the

OTHER condition (r = - 0.55, p < 0.01), and the correlation remained significant after we excluded the most extreme value, which could be considered as an outlier (r = - 0.43, p < 0.05, **Figure 3F**). That is, participants with greater other-regarding value signals in the VMPFC than in the DMPFC were more likely to choose to help their partners in the OTHER condition, whereas those with the opposite gradient were more likely to behave individualistically (selfishly). The spatial gradient for self-regarding values was not associated with the choices in the SELF condition (r = - 0.075).

Functional connectivity of the MPFC during other- vs. self-regarding choices. It has been well established by previous studies that communication between the medial frontal regions computing the chosen values and the striatum processing the RPEs plays an essential role in updating and maintaining value-related information during reinforcement learning (7, 8). We confirmed that activity in the striatum, including the nucleus accumbens and part of the caudate and putamen, was correlated with the RPE at the time of the outcome presentation phase, irrespective of an individual's propensity to behave prosocially (**SI Appendix**). We then performed psychophysiological interaction (PPI) analyses to test whether and how the individual differences observed in the present study are reflected in the pattern of functional connectivity between the MPFC subregions and the striatum during other vs. self-regarding choices. We selected the ventral (VMPFC; x = 0, y = 56, z = 2; peak voxel computing chosen value for SELF condition), middle (MMPFC; x = 4, y = 52, z = 8; peak voxel computing chosen value for both SELF and OTHER conditions), and dorsal (DMPFC; x = 2, y = 44, z = 12; peak voxel computing

chosen value for OTHER condition) parts of the MPFC as seed regions. Then we performed three separate PPI analyses to identify the regions showing differential functional coupling with the three seed regions in the OTHER vs. SELF condition at the option presentation phase. Finally, we performed two-sample t-tests of the difference between selfish and prosocial groups. As expected, we found a significant group difference in functional connectivity between the striatum and the VMPFC (x = 16, y = 20, z = 0, $Z = 3.54$), MMPFC (x = 6, y = 14, z = 2, $Z = 3.36$), and DMPFC (x = 14, y = 18, z = 4, $Z = 4.14$) (**Figures 4A** and **B**). To better understand these group differences, we performed post-hoc ROI analyses and found that the difference between the OTHER vs. SELF condition was robust among prosocial individuals, such that the connectivity was stronger in the OTHER condition than in the SELF condition for all of the three MPFC seeds [all Fs(1, 23) > 9.98, all ps < 0.01] (**Figures 4C-E**). On the other hand, selfish individuals tended to show the opposite patterns, that is, stronger connectivity in the SELF than the OTHER condition, although the differences between the two conditions were not statistically significant. It is also worth noting that the part of the striatum communicating with the DMPFC covered a large area that extended from the nucleus accumbens to the dorsal caudate and putamen. By contrast, the regions communicating with the MMPFC and VMPFC were more restricted, and the peak voxels were located more ventrally than those connected with the DMPFC (**SI Appendix, Figures 4A** and **B**; **Figure S5** and **Table S3** for the regions other than the striatum; **SI Appendix** for additional analyses addressing alternative explanations).

Mode of decisions for self and other. The difference between prosocial and selfish individuals in representing and updating self- and other-regarding values lead us to the question of whether the two groups engage in different modes of decision in SELF and OTHER conditions. Our response time (RT) data suggested the possibility that additional cognitive processes might be required for selfish individuals to make other-regarding decisions, because selfish individuals were significantly slower in the OTHER than the SELF condition, $F(2, 18) = 3.84$, $p < 0.05$ (**SI Appendix**, **Figure S8**). Prosocial individuals did not show such a difference, $F(2, 28) = 0.49$, ns. In line with this behavioral finding, regions known to be involved in cognitive control, such as the right anterior insula extending to the inferior frontal gyrus showed greater activation when selfish participants made choices for the other participant than for themselves, whereas prosocial individuals showed no significant difference across conditions (AI/IFG; x = 36, y = 26, z = -4, Z = 3.65; two-sample t-test for the group difference in the contrast maps of OTHER vs. SELF condition at the time of option presentation; **SI Appendix**, **Figure S9A**; see **Table S4** for whole-brain result). This result remained the same even when trial-to-trial RTs were included in the analysis (**SI Appendix**, **Figure S10**), ruling out the possibility that the increased AI/IFG activation during the OTHER condition among selfish participants may merely reflect differences in RT. Interestingly, the AI/IFG activation was correlated with the average RTs, such that participants with slower RTs in the OTHER vs. SELF condition showed greater AI/IFG activation in the OTHER vs. SELF condition ($r = 0.50$, $p < 0.05$), which remained significant even after controlling for

the effect of need for cognition (55). In sum, these findings suggest that the AI/IFG may be involved in decision mode switching, which is then indirectly related to additional information processing. The RTs and neural responses in the SELF condition were not related to performance.

To further examine how AI/IFG activation influences the corticostriatal communication underlying the process of updating and representing other- vs. self-regarding values, we correlated the AI/IFG activation (i.e., the beta estimates of the OTHER vs. SELF contrasts at the option presentation phase) with the MPFC-striatum connectivity (i.e., the beta estimates of the PPIs with the MPFC subregions as seeds during OTHER vs. SELF trials) across participants. Interestingly, the greater AI/IFG activation in the OTHER than the SELF condition, the weaker DMPFC-striatum (r = -0.48, p = 0.01), MMPFC-striatum (r = -0.47, p = 0.01), and VMPFC-striatum (r = -0.37, p = 0.11) coupling in the OTHER vs. SELF condition (**SI Appendix**, **Figure S9B**).

Discussion

The present study investigated the neural mechanisms of valuing and representing another's welfare and their relation to an individual's propensity for prosocial behavior. Combined with a computational approach, our prosocial learning task provided a novel behavioral measure to quantify individual differences in prosociality and allowed us to explore the question that we raised in the beginning: What makes some people more prosocial than others and how does our brain enable us to value the welfare of others?

Our finding that the spatial specificity for self vs. other-regarding value representation within the MPFC was robust only among selfish individuals and was attenuated among prosocial individuals supports the idea that prosociality requires a shared value representation for self and other (1, 17, 26). Interestingly, closer examination of the spatial gradient revealed that the difference between prosocial and selfish groups was especially prominent in the VMPFC. This might seem puzzling, considering that the VMPFC has been strongly implicated in the processes of self-relevant information (31-33) and the DMPFC in theory of mind and mentalizing (34-36). However, there is converging evidence for functional specialization within the MPFC. For example, the VMPFC has been suggested as a domain-general valuation system that processes significant and motivating information such as reward (37, 38), and the DMPFC as a part of the attentional system that is predominantly involved in cognitively demanding tasks such as strategic social inference (37-39). This idea does not necessarily contradict the idea that self-relevance is a major factor that distinguishes VMPFC and DMPFC (40) because self-relevant information is often most significant and motivating (19, 41). Our results suggest that the VMPFC is tightly associated with subjective valuation regardless of the choice's beneficiary (30), whereas the DMPFC is involved in more general other-specific processing commonly required for other-regarding choices invariant across individuals (22-24, 39, 42, 43).

Another interesting finding is that the pattern of functional coupling between the MPFC subregions computing the values of choices and the striatum encoding RPEs was

significantly correlated with individual propensity to help others. In support of this finding, many previous studies have reported strong anatomical and functional links between MPFC subregions and the striatum, which play a key role in reinforcement learning (7, 8, 44). More specifically, recent theoretical work proposed a hierarchical model, in which reinforcement learning in vertebrates occurs through multiple independent cortico-basal ganglia loops that interact with each other, allowing information to propagate mostly from ventral to dorsal levels of the striatum-MPFC loops (44). Although the spatial resolution of fMRI is far lower than that of animal neurophysiological studies (44), we found a similar spatial segregation between ventral and dorsal corticostriatal networks. Our PPI data suggest that prosocial individuals may be characterized by active propagation of subjective value signals between the ventral and dorsal loops, which could be crucial for maximizing their capacity to represent, update, and maintain the value of other-regarding choices. This, in turn, may enable the shared value representation for self and other within the MPFC.

Despite the stronger functional coupling between the striatum and the MPFC among prosocial individuals, it appears that selfish individuals required greater cognitive effort and control during the OTHER vs. SELF condition, where they had slower RTs and showed increased activity in the AI/IFG. Although we cannot completely rule out the possibility that the AI/IFG activation may reflect an aversive response to other-regarding choices among selfish individuals, the correlation between the AI/IFG activation and the average RTs across individuals suggests that additional cognitive processes may be

engaged during the OTHER condition. It is also noteworthy that AI/IFG has been strongly implicated in cognitive control and self-regulation (45-48). Given that increase in AI/IFG activity weakened the MPFC-striatum coupling during choices for others, selfish individuals seem to employ additional cognitively demanding processes that interrupt the process of prosocial valuation, which may involve signal propagation through the MPFC-striatum loops. The exact nature of these additional cognitive processes employed by selfish individuals deserves future investigation.

In summary, the present study revealed that spatial segregation within the MPFC in computing values for self- vs. other-regarding choice is critically involved in determining individual variability in prosociality. Further, weaker segregation of self- and other-regarding value signals in the MPFC and stronger MPFC-striatal coupling are associated with being prosocial rather than selfish. Despite having yet to be tested with more direct measures of altruism, our findings provide important insights into human prosociality/altruism. First, the shared neural representation for self- and other-regarding values found among prosocial individuals supports the view that altruism requires value extension from self to others, via a process in which another person's welfare becomes valuable (1, 16, 17, 26). Second, the other-regarding valuation process subserved by the VMPFC as a part of the corticostriatal network in prosocial individuals emphasizes the automatic and intuitive nature of prosocial motivation. This finding is in line with recent perspectives that prosociality and morality are rooted in intuition acquired, formed, and maintained through socialization (49-52). Third, our findings provide neural evidence of

social norms internalized within an individual, which may have evolved to benefit groups by promoting prosocial behaviors (53). In conclusion, our present findings shed some light on the mystery of human altruism and support the notion that this mystery can be better understood by adopting rigorous scientific methods and theoretical frameworks in the ripening field of decision neuroscience.

Materials and Methods

Participants. Thirty pairs of healthy right-handed female college students participated in the experiment. The two participants in each pair were strangers to each other. Half of the participants (one participant/pair; mean age: 21.9 years, range 19-29 years) were randomly assigned to perform a prosocial learning task in the scanner. Four participants were excluded due to excessive head movement or random responses in all the conditions, leaving twenty-six participants included in the fMRI analysis. All participants were compensated with 30,000KRW (approximately 30 USD). The study protocol was approved by the Korea University Institutional Review Board and all participants gave written consent to participate before the experiment.

Prosocial learning task. During the prosocial learning task, participants made choices between two fractal images, each of which was associated with different reward probabilities (30% vs. 70%) (see **SI Appendix**). Each trial began with one of three pairs of fractal images and each pair of images was associated with one of three different types of condition: SELF, BOTH, or OTHER (**Figure 1**). Participants could earn two points for

self and none for other in the SELF condition, one point for self and one point for other in the BOTH condition, and none for self and two points for other in the OTHER condition. The points earned from ten randomly selected trials across all three conditions were to be used to reduce the duration of exposure to stressful noise (i.e., 10 sec per point) for self and/or other. Each condition comprised 48 trials, resulting in a total of 144 trials in one functional run (about 25 min). The conditions were presented in a pseudo-random order, and the conditions and reward probabilities associated with different fractal images were counter-balanced across participants.

Estimation of chosen value and reward prediction error. Chosen value and reward prediction error (RPE) for each trial for each individual were estimated by using the advantage learning model (3, 4) (see **SI Appendix**).

Behavioral measures of individual prosociality. To measure individual differences in prosociality within the task, we estimated an experienced magnitude of reward outcomes separately for each of the three conditions (see **SI Appendix**). Applying the conceptual framework of social value orientation (54), we grouped participants into a prosocial group ($N = 15$), which valued other-regarding outcomes the same as or more than self-regarding outcomes, and a selfish group ($N = 10$), which valued self-regarding outcomes more than other-regarding outcomes. One subject who did not value either of the outcomes was excluded from the analyses examining group differences. To validate the findings that were based on the task-based measure of prosociality with an independent measure, we also used participants' self reports in the social value orientation

questionnaire to group them into prosocial (N = 14) and proself (N = 10) groups (53; see **SI Appendix**). This grouping confirmed that the behavioral and fMRI results remained the same (**SI Appendix**, **Figures S2** and **S4**).

fMRI data acquisition and analysis. Brain images were acquired on a 3T MRI scanner (MAGNETOM Tim Trio; Siemens Medical Solutions, Erlangen, Germany) at the Korea University Brain Imaging Center. T2*-weighted functional images were obtained through gradient echo planar image (EPI) with BOLD (Blood Oxygenation Level Dependent) contrast (TR = 2000 ms; TE = 30 ms; flip angle = 90°; FOV = 240mm; 80 × 80 matrix; 36 axial slices; 3 × 3 × 3 mm in-plane resolution). High-resolution T1-weighted structural images were also collected (TR = 1900 ms; TE = 2.52 ms; flip angle = 9°; 256 × 256 matrix; 1 × 1 × 1 mm in-plane resolution). The fMRI data were preprocessed and analyzed using SPM8 (Wellcome Department of Imaging Neuroscience, London). Images were realigned, normalized to the standard Montreal Neurological Institute (MNI) EPI template, and spatially smoothed using a Gaussian kernel with an 8-mm full width at half maximum (FWHM).

We created a first-level general linear model (GLM) with parametric modulators (see **SI Appendix**). Trial-by-trial fluctuations of subject-specific chosen values and RPEs were estimated by using the advantage learning model and entered into the first-level GLM model as parameters that modulated the hemodynamic responses at the time of option presentation and outcome presentation, respectively. Linear contrasts of regression coefficients for the parametric modulators of value and for RPE were computed and

subjected to a random effects group-level analysis using one-way ANOVA with condition (i.e., SELF, BOTH, OTHER) as a repeated-measures factor. Spatial gradients in self- and other-regarding value representations within the MPFC were quantified by extracting parameter estimates from five anatomical ROIs (spheres with a 4-mm radius) along the midline axis from VMPFC to DMPFC within the activation cluster correlating with the value parameters (30). For each ROI, we extracted the value parameter estimates for each condition, which were then entered into a repeated-measures one-way ANOVA. In addition, we fitted a linear slope to the OTHER vs. SELF contrasts across the five ROIs along the ventral-to-dorsal axis for each individual to estimate the degree of spatial gradient within the MPFC in terms of other- vs. self-regarding valuation.

Differential functional connectivity with the MPFC subregions during other- compared to self-regarding choices at option presentation was assessed with a PPI analysis. The MMPFC seed was the peak voxel from the region ($x = 4$, $y = 52$, $z = 8$) that was found to be correlated with the value parameters from all three conditions in the parametric modulation analysis; the VMPFC ($x = 0$, $y = 56$, $z = 2$) and DMPFC ($x = 2$, $y = 44$, $z = 12$) seeds were the peak voxels found to be correlated with the value parameters from SELF and OTHER conditions, respectively; the OTHER vs. SELF contrast at the time of option presentation was included as a psychological variable. Individual PPI maps were entered into a group-level two-sample t-test comparing prosocial and selfish groups.

Additionally, to examine whether prosocial and selfish individuals utilize different modes of decision for self vs. other, we contrasted neural responses to the presentation of options between SELF and OTHER conditions. The contrast maps of SELF vs. OTHER and OTHER vs. SELF at option presentation were entered into random-effects group-level two-sample t-tests comparing prosocial and selfish groups.

All statistical thresholds were set to $p < 0.05$ corrected for multiple comparisons, using a cluster threshold determined at an uncorrected $p < 0.001$ by Monte Carlo simulations implemented in AlphaSim within AFNI software (http://afni.nih.gov/afni; 55) for each search volume described below. To assess value signals in the MPFC, we formed an *a priori* anatomical search volume that included superior medial frontal cortex and anterior cingulate based on the AAL atlas (57) as implemented in the WFU_PickAtlas toolbox (www.ansir.wfubmc.edu; 58). In searching for RPE signals in the striatum and for the PPI analyses, we created an *a priori* search volume including bilateral caudate and putamen (extending to nucleus accumbens), based on the AAL atlas. For the group comparison analyses of SELF vs. OTHER and OTHER vs. SELF contrasts at option presentation, the correction was confined within the whole-brain because we had no specific hypothesis for this analysis.

Author Contributions. SS, PNT, SL, GH, EF, and HK conceived and designed the experiments. SS and DJ performed the experiments. SS and HK analyzed the data. SS and HK contributed reagents/materials/analysis tools. SS, PNT, and HK wrote the paper.

Acknowledgments. This work was supported by a research grant from Korea University (SS), a National Research Foundation of Korea Grant from the Korean Government (NRF-2012-S1A3-A2033375: SS and HK; 2006-2005110: HK) and the Swiss National Science Foundation (PP00P1_128574 and PP00P1_150739: PNT; CRSII3_141965: EF and PNT)

References

1. Batson CD (2011) *Altruism in humans* (Oxford University Press, New York).
2. Thorndike EL (1911) *Animal intelligence*: Experimental studies (Macmillan, New York).
3. Kim H, Shimojo S, O'Doherty JP (2006) Is avoiding an aversive outcome rewarding? Neural substrates of avoidance learning in the human brain. *Plos Biol* 4:e233.
4. O'Doherty JP, Dayan P, Schultz J, Deichmann R, Friston K, Dolan, RJ (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* 304:452-454.
5. Sutton RS, Barto AG (1998) *Introduction to reinforcement learning* (MIT Press, Cambridge, MA).
6. Hare TA, O'Doherty J, Camerer CF, Schultz W, Rangel A (2008) Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *J Neurosci* 28:5623-5630.
7. Haber SN, Knutson B (2009) The reward circuit: linking primate anatomy and human imaging. *Neuropsychopharmacology* 35:4-26.
8. Haruno M, Kawato M (2006) Heterarchical reinforcement-learning model for integration of multiple cortico-striatal loops: fMRI examination in stimulus-action-reward association learning. *Neural Networks* 19:1242-1254.
9. Kim, H, Adolphs, R, O'Doherty, JP, and Shimojo, S (2007). Temporal isolation of neural processes underlying face preference decisions. *Proc Natl Acad Sci USA,* 104:18253-18258.
10. Schultz W (1998) Predictive reward signal of dopamine neurons. *J Neurophysiol* 80:1-27.

11. Lebreton M, Jorge S, Michel V, Thirion B, Pessiglione M (2009) An automatic valuation system in the human brain: evidence from functional neuroimaging. *Neuron* 64:431-439.
12. Plassmann H, O'Doherty J, Rangel A (2007) Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. *J Neurosci* 27:9984-9988.
13. Behrens TE, Hunt LT, Rushworth MFS (2009) The computation of social behavior. *Science* 324:1160-1164.
14. Burke CJ, Tobler PN, Baddeley M, Schultz W (2010) Neural mechanisms of observational learning. *Proc Natl Acad Sci USA* 107:14431-14436.
15. Christopoulos GI, King-Casas B (in press) With you or against you: Social orientation dependent learning signals guide actions made for others. *NeuroImage*.
16. Fehr E, Camerer CF (2007) Social neuroeconomics: the neural circuitry of social preferences. *Trends Cogn Sci* 11:419-427.
17. Harbaugh WT, Mayr U, Burghart DR (2007) Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science* 316: 1622-1625.
18. Izuma K, Saito DN, Sadato N (2008) Processing of social and monetary rewards in the human striatum. *Neuron* 58:284-294.
19. Seid-Fatemi A, Tobler PN (in press) Efficient learning mechanisms hold in the social domain and are implemented in the medial prefrontal cortex. *Soc Cogn Affect Neurosci.*
20. Chang SWC, Gariepy JF, Platt ML (2013) Neuronal reference frames for social decisions in primate frontal cortex. *Nat Neurosci* 16:243-250.
21. Denny BT, Kober H, Wager TD, Ochsner KN (2012) A Meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. *J Cognitive Neurosci* 24:1742-1752.
22. Mitchell JP, Macrae CN, Banaji MR (2006) Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron* 50:655-663.
23. Moran JM, Lee SM, Gabrieli JDE (2010) Dissociable neural systems supporting knowledge about human character and appearance in ourselves and others. *J Cognitive Neurosci* 23:2222-2230.
24. Qin P, Northoff G (2011) How is our self related to midline regions and the default-mode network? *NeuroImage* 57:1221-1233.
25. Saxe R, Moran JM, Scholz J, Gabrieli J (2006) Overlapping and non-overlapping brain regions for theory of mind and self reflection in individual subjects. *Soc Cogn Affect Neurosci* 1:229-234.
26. Hare TA, Camerer CF, Knoepfle DT, O'Doherty JP, Rangel A (2010) Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. *J Neurosci* 30:583-590.

27. Moll J, Krueger F, Zahn R, Pardini M, De Oliveira-Souza R, Grafman J (2006) Human fronto-mesolimbic networks guide decisions about charitable donation. *Proc Natl Acad Sci USA* 103:15623-15628.
28. Shenhav A, Greene JD (2010) Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron* 67:667-677.
29. Tricomi E, Rangel A, Camerer CF, O'Doherty JP (2010) Neural evidence for inequality-averse social preferences. *Nature* 463:1089-1091.
30. Nicolle A, Klein-Flügge MC, Hunt LT, Vlaev I, Dolan RJ, Behrens TE (2012) An agent independent axis for executed and modeled choice in medial prefrontal cortex. *Neuron* 75:1114-1121.
31. Heatherton TF, Wyland CL, Macrae CN, Demos KE, Denny BT, Kelley WM (2006) Medial prefrontal activity differentiates self from close others. *Soc Cogn Affect Neurosci* 1:18-25.
32. Northoff G, Heinzel A, de Greck M, Bermpohl F, Dobrowolny H, Panksepp J (2006) Self-referential processing in our brain—A meta-analysis of imaging studies on the self. *NeuroImage* 31:440-457.
33. Sul S, Choi I, Kang P (2012) Cultural modulation of self-referential brain activity for personality traits and social identities. *Soc Neurosci* 7:280-291.
34. Frith CD, Frith U (1999) Interacting minds--a biological basis. *Science* 286:1692-1695.
35. Hampton AN, Bossaerts P, O'Doherty JP (2008) Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proc Natl Acad Sci USA* 105:6741-6746.
36. Saxe, R (2006) Uniquely human social cognition. *Curr Opin Neurobiol* 16:235-239.
37. Bartra O, McGuire JT, and Kable JW (2013) The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage* 76:412-427.
38. Bzdok D, Langner R, Schilbach L, Engemann DA, Laird AR, Fox PT, Eickhoff S (2013) Segregation of the human medial prefrontal cortex in social cognition. *Front in Hum Neurosci 7*.
39. Seo H, Cai X, Donahue CH, Lee D (2014) Neural correlates of strategic reasoning during competitive games. *Science* 346:340-343.
40. Wagner DD, Haxby JV, Heatherton TF (2012) The representation of self and person knowledge in the medial prefrontal cortex. *Wiley Interdiscip Rev Cogn Sci* 3:451-470.
41. Verplanken B, Holland RW (2002) Motivated decision making: Effects of activation and self-centrality of values on choices and behavior. *J Pers Soc Psychol* 82:434-447.

42. Jung D, Sul S, and Kim H (2013) Dissociable neural processes underlying risky decisions for self versus other. *Front in Neurosci* 7:15. doi:10.3389/fnins.2013.00015
43. Kang P, Lee J, Sul S, Kim H (2013) Dorsomedial prefrontal cortex activity predicts the accuracy in estimating others' preferences. *Front in Hum Neurosci* 7:686. doi:10.3389/fnhum.2013.00686.
44. Yin HH, Knowlton BJ (2006) The role of the basal ganglia in habit formation. *Nat Rev Neurosci* 7:464-476.
45. Aron AR, Durston S, Eagle DM, Logan GD, Stinear CM, Stuphorn V (2007) Converging evidence for a fronto-basal-ganglia network for inhibitory control of action and cognition. *J Neurosci* 27:11860-11864.
46. Corbetta M, Shulman GL (2002) Control of goal-directed and stimulus-driven attention in the brain. *Nat Rev Neurosci* 3:201-215.
47. Heatherton TF (2011) Neuroscience of self and self-regulation. *Annu Rev Psychol* 62:363-390.
48. Sridharan D, Levitin DJ, Menon V (2008) A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. *Proc Natl Acad Sci USA* 105:12569-12574.
49. Greene JD, Paxton JM (2009) Patterns of neural activity associated with honest and dishonest moral decisions. *Proc Natl Acad Sci USA* 106:12506-12511.
50. Rand DG, Greene JD, Nowak MA (2012) Spontaneous giving and calculated greed. *Nature* 489:427-430.
51. Haidt J (2007) The new synthesis in moral psychology. *Science* 316:998-1002.
52. Sauer H (2012) Educated intuitions. Automaticity and rationality in moral judgement. *Philos Explor* 15:255-275.
53. Sober E, Wilson DS (1998) *Unto Others: the evolution and psychology of unselfish behavior* (Harvard University Press, Cambridge, MA).
54. Van Lange PAM, De Bruin EMN, Otten W, Joireman JA (1997) Development of prosocial, individualistic, and competitive orientations: Theory and preliminary evidence. *J Pers Soc Psychol* 73:733-746.
55. Cacioppo JT, Petty RE, Feng Kao C (1984) The efficient assessment of need for cognition. *J Pers Assess* 48:306-307.
56. Cox RW (1996) AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput Biomed Res* 29:162-173.
57. Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Joliot M (2002) Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage* 15:273-289.
58. Maldjian JA, Laurienti PJ, Kraft RA, Burdette JH (2003) An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *NeuroImage* 19:1233-1239.

Figure Legends

Figure 1 Prosocial learning task and within-subject experimental conditions. Only rewarded outcomes are shown.

Figure 2 Behavioral results. A: Proportions of choosing high reward probability (HRP) options in SELF, BOTH, and OTHER conditions. B: Comparison between prosocial and selfish. All error bars indicate 95% confidence interval.

Figure 3 Regions representing the value of the chosen option. A: Definition of ROIs. B: Spatial gradient for self- and other- regarding value computation within the MPFC. Dotted lines indicate linear fits of the spatial gradient for SELF (red) and OTHER (blue) conditions. C-D: Spatial gradient for self- and other- regarding value computation within the MPFC, depicted separately for prosocial (N = 15) and selfish groups (N = 10) as defined by the advantage learning model (see **SI Appendix**). Dotted lines indicate linear fittings of the spatial gradient for SELF (red) and OTHER (blue) conditions. Negative slope indicates greater value signal in VMPFC than DMPFC, and *vice versa*. E: Linear slopes of the spatial gradient within the MPFC in SELF and OTHER conditions among prosocial and selfish participants. Negative slope indicates greater value signal in

VMPFC than DMPFC, and *vice versa*. F: Participants with greater gradient showed a less clear preference for the high probability option in the OTHER condition. The result remained significant after excluding the subject with the strongest gradient (open circle, correlation coefficient without this data point is reported in brackets).

Figure 4 Comparison of prosocial vs. selfish individuals for functional connectivity (PPI) with the VMPFC, MMPFC, and DMPFC as seed regions during OTHER vs. SELF conditions. A: Result of two-sample t-tests for the PPIs with the VMPFC and DMPFC as seed regions masked with the striatum ROI. Different seed regions are color coded (Red: VMPFC, blue: DMPFC, magenta: VMPFC $\cap$ DMPFC; for illustration purposes, $p <$ 0.005, uncorrected). B: Result of two-sample t-tests for all three PPIs masked with the striatum ROI. Voxels connected with the VMPFC, MMPFC, and DMPFC as seed regions are color coded with red, green, and blue, respectively (note that the regions connected with VMPFC and MPFC largely overlap; for illustration purposes, $p < 0.005$, uncorrected). C-E: Average connectivity strength between the MPFC subregions and the striatum in SELF and OTHER condition closely matched the differential prosociality of the prosocial and the selfish group.